Part-of-speech Tagging

Yuanjing Zhu

October 11, 2022

1 Introduction

In PosTagging.py, I built a part-of-speech hidden markov model using the first 10,000 tagged sentences from the Brown corpus to infer the sequence of states for sentences. First, I generated 3 matrix: initial matrix, transition matrix, and observation matrix. Then I used the provided Viterbi implementation with an OOV observation and smoothing everywhere to test the function on first 10150-10152 sentences from the Brown corpus.

The three sentences are :

1. Those coming from other denominations will welcome the opportunity to become informed.
2. The preparatory class is an introductory face-to-face group in which new members become aquainted with one another.
3. It provides a natural transition into the life of the local church and its organizations.

2 Result

Comparing the result from my implementation against the truth, the accuracy is 91.5%. There are 47 words in total, of which the model could correctly tag 43.

Sentence	POS sequence
S1	['Those', 'coming', 'from', 'other', 'denominations', 'will', 'welcome', 'the', 'opportunity', 'to', 'become', 'informed', '.']
S1_Truth	['DET', 'VERB', 'ADP', 'ADJ', 'NOUN', 'VERB', 'VERB', 'DET', 'NOUN', 'PRT', 'VERB', 'VERB', '.']
S1_Output	['DET', 'NOUN', 'ADP', 'ADJ', 'NOUN', 'VERB', 'VERB', 'DET', 'NOUN', 'PRT', 'VERB', 'VERB', '.']
S2	['The', 'preparatory', 'class', 'is', 'an', 'introductory', 'face-to-face', 'group', 'in', 'which', 'new', 'members', 'become', 'acquainted', 'with', 'one', 'another', '.']
S2_Truth	['DET', 'ADJ', 'NOUN', 'VERB', 'DET', 'ADJ', 'ADJ', 'NOUN', 'ADP', 'DET', 'ADJ', 'NOUN', 'VERB', 'VERB', 'ADP', 'NUM', 'DET', '.']
S2_Output	['DET', 'ADJ', 'NOUN', 'VERB', 'DET', 'NOUN', 'ADP', 'NOUN', 'ADP', 'DET', 'ADJ', 'NOUN', 'VERB', 'VERB', 'ADP', 'NUM', 'NOUN', '.']
S3	['It', 'provides', 'a', 'natural', 'transition', 'into', 'the', 'life', 'of', 'the', 'local', 'church', 'and', 'its', 'organizations', '.']
S3_Truth	['PRON', 'VERB', 'DET', 'ADJ', 'NOUN', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'ADJ', 'NOUN', 'CONJ', 'DET', 'NOUN', '.']
S3_Output	['PRON', 'VERB', 'DET', 'ADJ', 'NOUN', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'ADJ', 'NOUN', 'CONJ', 'DET', 'NOUN', '.']

	.	ADP	PRON	VERB	X	PRT	CONJ	NUM	DET	ADJ	NOUN	ADV
.	0.141609	0.111374	0.063895	0.130743	0.002008	0.020196	0.093835	0.021377	0.120113	0.052852	0.178930	0.063068
ADP	0.011410	0.018955	0.049687	0.038535	0.000405	0.012219	0.001325	0.039050	0.442989	0.084579	0.287192	0.013655
PRON	0.081783	0.049618	0.009375	0.732151	0.000144	0.022501	0.011827	0.000865	0.014135	0.009231	0.009087	0.059282
VERB	0.071596	0.165493	0.039010	0.193569	0.000216	0.063442	0.012015	0.012602	0.176798	0.058377	0.111163	0.095719
X	0.255639	0.071429	0.003759	0.052632	0.443609	0.007519	0.022556	0.007519	0.007519	0.007519	0.105263	0.015038
PRT	0.046672	0.089011	0.003348	0.660299	0.000394	0.008862	0.008074	0.007286	0.084679	0.019102	0.041552	0.030721
CONJ	0.023077	0.067722	0.052640	0.160181	0.000452	0.023680	0.000603	0.019608	0.157014	0.117496	0.289442	0.088084
NUM	0.253436	0.140034	0.005441	0.044101	0.000573	0.006873	0.033792	0.023482	0.010596	0.068442	0.381157	0.032073
DET	0.012933	0.009038	0.008414	0.061473	0.001558	0.001597	0.000545	0.013401	0.006467	0.250292	0.615388	0.018894
ADJ	0.089420	0.079734	0.002436	0.015533	0.000670	0.018578	0.033319	0.012853	0.006091	0.059268	0.674423	0.007675
NOUN	0.264596	0.228648	0.017747	0.144012	0.000487	0.017478	0.052519	0.009714	0.014117	0.015680	0.212178	0.022823
ADV	0.139848	0.138260	0.036312	0.258946	0.000318	0.029430	0.013762	0.015562	0.080669	0.150328	0.040229	0.096337

Figure 1: Transition matrix

The POS tagger could generate majority of correct tags because the training samples are pretty large, total 10,000 sentences, and the 3 test sentences come from the same corpus of the training data, so the sentence structure and the frequently used words are similar.

For the 4 mis-tagged words, the first word is "coming", it is a "VERB" but I tagged it as "NOUN". From the transition matix, we can see that the probability of a NOUN following a DET is 61.5% while that of a VERB following a DET is only 6.14%, so the model tends to choose NOUN to maximize the probability. For the second mis-tagged word "introductory", the reason why the model made a mistake is similar, the probability of an ADJ following a DET is 25.0%, which is lower than the DET-NOUN combination. Next, the model failed to tag "face-to-face" as "ADJ" because it is an unknown word in the training set, so the model just identified it as "ADP" for its 22.9% likelihood. Finally, the model mistakenly tagged "another" as NOUN which should be a DET because the probability of a DET following a NUM is pretty low, only 1.06%.